Published online 5 May 2009 *Nucleic Acids Research, 2009, Vol. 37, Web Server issue* ***W229–W234***
doi:10.1093/nar/gkp286

Insignia: a DNA signature search web server for diagnostic assay development

Adam M. Phillippy[1,*], Kunmi Ayanbule[1], Nathan J. Edwards[1,2] and Steven L. Salzberg[1]

[1]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD and [2]Department of Biochemistry and Molecular & Cellular Biology, Georgetown University Medical Center, Washington, DC, USA

Received January 29, 2009; Revised April 3, 2009; Accepted April 14, 2009

ABSTRACT

Insignia is a web application for the rapid identification of unique DNA signatures. DNA signatures are distinct nucleotide sequences that can be used to detect the presence of certain organisms and to distinguish those organisms from all other species. These signatures can be used as the basis for diagnostic assays to detect and genotype microbes in both environmental and clinical samples. Insignia identifies an exhaustive set of accurate DNA signatures for any set of target genomes, and screens these signatures against a comprehensive background that includes all sequenced bacteria and viruses, the human genome, and many other animals and plants. Identified signatures may be browsed by genomic location or proximal genes, filtered by composition, viewed in a genome browser or directly downloaded. Integrated PCR primer design is also provided for each signature. The Insignia website (http://insignia.cbcb.umd.edu) is free and open to all users and there is no login requirement. In addition, the source code for the computational pipeline is freely available.

INTRODUCTION

Insignia provides a convenient web interface for identifying genomic signatures from a database of all current bacterial and viral genomic sequences, which currently comprises over 8300 distinct organisms. The input to Insignia is any set of target and background genomes selected from the online database, and the output is a list of signatures perfectly conserved by all target genomes and absent from all of the background genomes. Insignia is the only web application capable of performing this whole-genome signature design, whereby entire genomes are screened against a comprehensive background. KPATH (1) and TOFI (2) perform similar computations, but must be run offline and require considerable computational resources. To quickly identify signatures for any combination of target and background genomes, and enable use over the web, Insignia maintains a specialized database containing pre-computed matches between every pair of genomes. Using this match information, signatures are computed 'on the fly' in a matter of seconds, using efficient interval set operations. A full description of Insignia's computational algorithms is given in Phillippy *et al*. (3). The following sections describe the associated web application, which provides an easy interface for signature search and assay design.

OVERVIEW

A k-mer signature is a string of k nucleotides that is perfectly conserved in a set of target genomes, but does not occur exactly in any of the background genomes. Given a set of target genomes, a set of background genomes, and a signature length k, Insignia identifies all k-mer signatures present in the target genomes. A short k-mer (e.g. 15–20 bp) not only is more likely to be shared by a group of target species, but is also more likely to appear in the background. When Insignia finds a series of overlapping k-mer signatures, it reports these longer chains as a single region, where every k-mer in the chain is guaranteed to be unique. Occasional background k-mer matches will occur by chance, but long chains of k-mer signatures are likely to represent the sequences most dissimilar from the background. This can be thought of as the inverse of typical seed-and-extend alignment strategies such as BLAST. Instead of assuming similarity exists in regions sharing exact matches, users of Insignia can assume that dissimilarity exists in regions devoid of exact matches.

Signature chains are reported as an interval from the start of the first signature word to the end of the last signature word in the chain. The signature chain $[s,e]$ contains exactly $(e - s - k + 2)$ signature words of length k, completely covering the interval $[s,e]$ in the target sequence. Signature words are, by definition, perfectly conserved in all target genomes, and contain at least a

*To whom correspondence should be addressed. Tel: +1 301 405 3234; Fax: +1 301 314 1341; Email: amp@umiacs.umd.edu

single difference from every background sequence. (Note that a signature may occur multiple times in a target genome; it is not required to occur just once.) Therefore, a signature chain will contain a difference from any background genome at least every k bases. For some types of detection assays, a difference every ~20 bp is not sufficient for discrimination. However, polymorphisms tend to be unevenly distributed, and similar sequences are likely to share at least one exact match over a long distance. In our validation studies, we have found that long signature chains (e.g. >100 bp) follow this tendency and are often quite dissimilar from the background. After identifying candidate signatures, a more sensitive BLAST search of the background can be performed to identify any similar but non-identical matching sequences.

Signature chains have the benefit of being both long and specific. Thus, they make ideal targets for PCR-based detection assays such as TaqMan, which work well with an amplicon length of around 100 bp and specific primers and probes. For microarrays, signature chains can be tiled across their length with multiple probes to provide adequate redundancy. These two techniques can be combined to achieve a very high degree of accuracy. PCR primers can be designed at the boundaries of a long signature chain, and the interior of the chain can be tiled with microarray probes. The detection procedure could then consist of PCR amplification, followed by microarray hybridization or sequencing of the product. Alternative probe-based detection strategies, such as melting curve analysis (4), are also available that are sensitive to a single nucleotide difference in the probe sequence.

IMPLEMENTATION

The Insignia software is organized into two primary components: the web interface and the computational pipeline. The computational pipeline is a standalone component, written in C++, that generates match data offline on a computational grid and stores the results in indexed files. Signature queries can be transmitted to the pipeline, which retrieves and processes match data from the index files to compute signatures. The dynamic web interface submits queries to the computational pipeline and displays the results. The web interface is written in HTML, JavaScript, PHP and Perl. The signature, sequence and annotation data used by the web interface are stored in a mySQL database using a Chado schema (5). Sybil provides web displays of sequence and annotation information (http://sybil.sourceforge.net).

INTERFACE

The Insignia home page provides a full listing of all target genomes currently in the database, organized alphabetically. The database mirrors the Gemina database, maintained by the Institute for Genome Sciences (6). Currently, the Gemina database includes 8341 distinct organisms, comprising 5506 viruses and 2835 bacteria. This list includes multiple genomes for many well-known species. For instance, only a single *Vibrio cholerae* genome sequence existed in 2006, when the original validation study was performed, but the database now contains 17 *V. cholerae* genomes. The database is updated routinely, and the accuracy of the predicted signatures will continue to increase as more genomes are added.

General help information and examples are also linked from the home page. Links are provided to automatically run signature searches for some important pathogens. Additional links are provided in the 'Recent Searches' box after custom searches are performed. This allows users to rerun a previous search with identical parameters. This is helpful for users who may return to the same results many times over the course of designing an assay.

In addition to displaying the current database status and search links, the Insignia home page also hosts a collection of validated signatures. These signatures were predicted by Insignia and have undergone laboratory validation against a series of samples including phylogenetic and environmental neighbors. Validated signatures are currently available for *V. cholerae, Francisella tularensis*, *Burkholderia mallei* and *Burkholderia pseudomallei*.

SEARCH PAGE

The Insignia Search page is available under the 'Run Insignia' link on the home page. On the Search page, users are asked to select a reference genome, a set of target genomes, a background and a signature word length. The reference genome is a member of the target set that provides the genome coordinates for purposes of reporting signatures, and that is used for displaying signatures along with genome annotation. Therefore, it is preferable to select a finished and annotated genome as the reference. The reference genome may be selected from a list of all genomes in the database, or users may type the name of the genome in the search box and the list will be automatically filtered for genomes matching that name. Additional target genomes may be selected from a list in a similar way as the reference genome, or selected from a taxonomy tree of closely related organisms. Figure 1 shows a partial view of the target tree for a *V. cholerae* reference. The tree view makes it easy to select all strains from a species or group of related species.

After a reference and target have been selected, users may modify the background set if necessary. By default, the background includes all genomes in the database, except for the genomes that have been selected as targets. Buttons are available to exclude draft genomes from the background, or to exclude draft genomes in the same genus as the reference. Draft genomes may contain large gaps, and including them in the target set may have the unintended consequence that no shared signatures can be found; for this reason, it is best to ignore draft genomes unless they are known to be nearly complete. We have found that target genomes with >100 contigs may interfere with signature computations and should probably be excluded. The interface provides an option to exclude all genomes with greater than a certain number of contigs

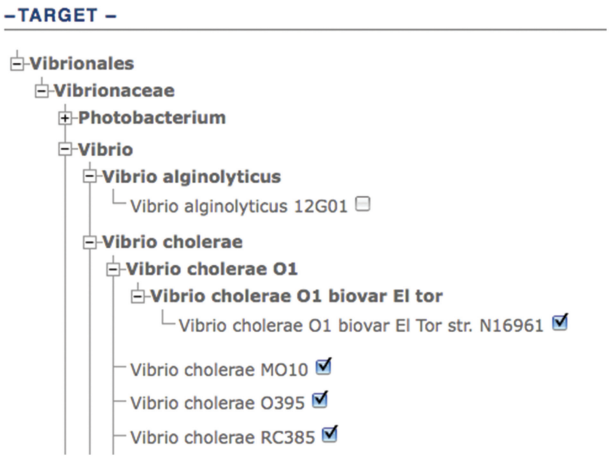

Figure 1. Taxonomic selection tree showing the order Vibrionales. Subtrees can be expanded or collapsed by clicking, and checkboxes are provided for adding specific genomes to the target set.

from both the target and background. The interface provides a simple check box to allow the user to exclude all draft genomes from the background.

Insignia is primarily designed for detecting bacterial signatures, but it can equally well design signatures for viruses. Viruses tend to have much smaller genomes and much higher mutation rates (especially RNA viruses), with the consequence that for a given viral species, there might not be any shared signatures among the sequenced isolates. However, if any signatures exist for a set of viruses, Insignia is guaranteed to find them. To simplify usage for bacterial assay design, all viral genomes can be hidden from view with a checkbox, but they will remain in the background unless specifically excluded.

The final step on the Search page is to select a signature word length. Insignia is capable of identifying k-mer signatures for any $k \geq 18$, but as the size of the background continues to increase, many k-mers will hit the background simply by chance and reduce the length of the resulting signature chains. A small k is desirable because it increases the minimum frequency of differences between a signature chain and the background. For the current database size, $k = 20$ appears to be a reasonable compromise between long signature chains and small k. We tested Insignia with $k = 20$ and a background including all of the NCBI RefSeq genomes (7), ~80 billion nucleotides, and signature chains >100 bp were still identified for most bacterial species. In future versions of Insignia, if the signature chains are insufficiently long for a PCR product, we plan to provide users the ability to design assays across multiple, neighboring signature chains.

Insignia also provides species-specific, e-unique signatures for most reference genomes. An e-unique signature is a k-mer signature that requires at least $(e + 1)$ string edits (substitutions, insertions, deletions) to match the background of other species. These pre-computed signatures are specific to the reference genome, but are not necessarily conserved among other genomes of the same species. If they coincide with an exact-match signature for that species, than they are both e-unique and conserved throughout the species. As with the exact-match signatures, adjacent e-unique signatures are displayed as chains. If e-unique signatures are available for the selected reference genome, a checkbox appears next to the selection box, giving users the option of including these signatures in the search results. 1- and 2-unique species-specific signatures for word lengths 18 and 19 bp are currently available for most finished bacterial genomes via the web interface, and 1-unique signatures for word lengths 20 through 25 are available from the authors upon request.

In addition to the Search page, Insignia accepts query submission via URLs to support links from external sites. The URL format contains the requested signature word length, and the GenBank taxonomy identifiers (TaxID) of the target genomes. The first TaxID is taken as the reference genome, and everything else as the target. If the first TaxID is an internal taxonomy branch, all children are included as the target and the user will be requested to select a reference. For example, to find 20-mer signatures for all strains in the species *Bacillus anthracis*, the URL is (http://insignia.cbcb.umd.edu/results.php?len = 20&taxid = 1392), where '1392' is the GenBank taxonomy identifier for the species *B. anthracis*. Additional TaxIDs can be added as a comma separated list.

The Gemina website uses this URL interface to link to Insignia (http://gemina.igs.umaryland.edu). Gemina is a database of epidemiology metadata linked to the genomes of infectious pathogens. Through this website, users can search for pathogens via their associated metadata such as transmission method, hosts, symptoms or geographical location. After a group of organisms with the desired traits are selected on the Gemina website, an Insignia search can be launched to identify genomic signatures for those organisms.

RESULTS PAGE

Figure 2 shows the Results page of the Insignia website after a signature search has completed. The target of this signature search was all 17 strains of *V. cholerae*, with all other genomes as the background. At this stage, all signatures can be downloaded in bulk for further analysis, or browsed on the Results page. The center of the page displays the resulting signature chains (hereafter referred to only as 'signatures'). By default, a table of signatures is displayed with start and stop positions, and the signature sequence. The left of the page provides links to relevant websites, help information and recent searches. The right of the page provides dynamic JavaScript controls for filtering the signatures. In the page displayed in Figure 2, 55 385 signatures have been identified but the filter has been set to show only those ≥100 bp in length, which left 1503 for display. Using the bottom two checkboxes, annotation information was added to the table and the signatures were sorted in order of decreasing length. In addition, the graphical display has been enabled by choosing 'View Graphical Output' from the Submit menu in the center of the page.

The JavaScript genome browser in the center of the page supports dynamic selection, zooming and panning, and enables users to browse the selected signatures in the

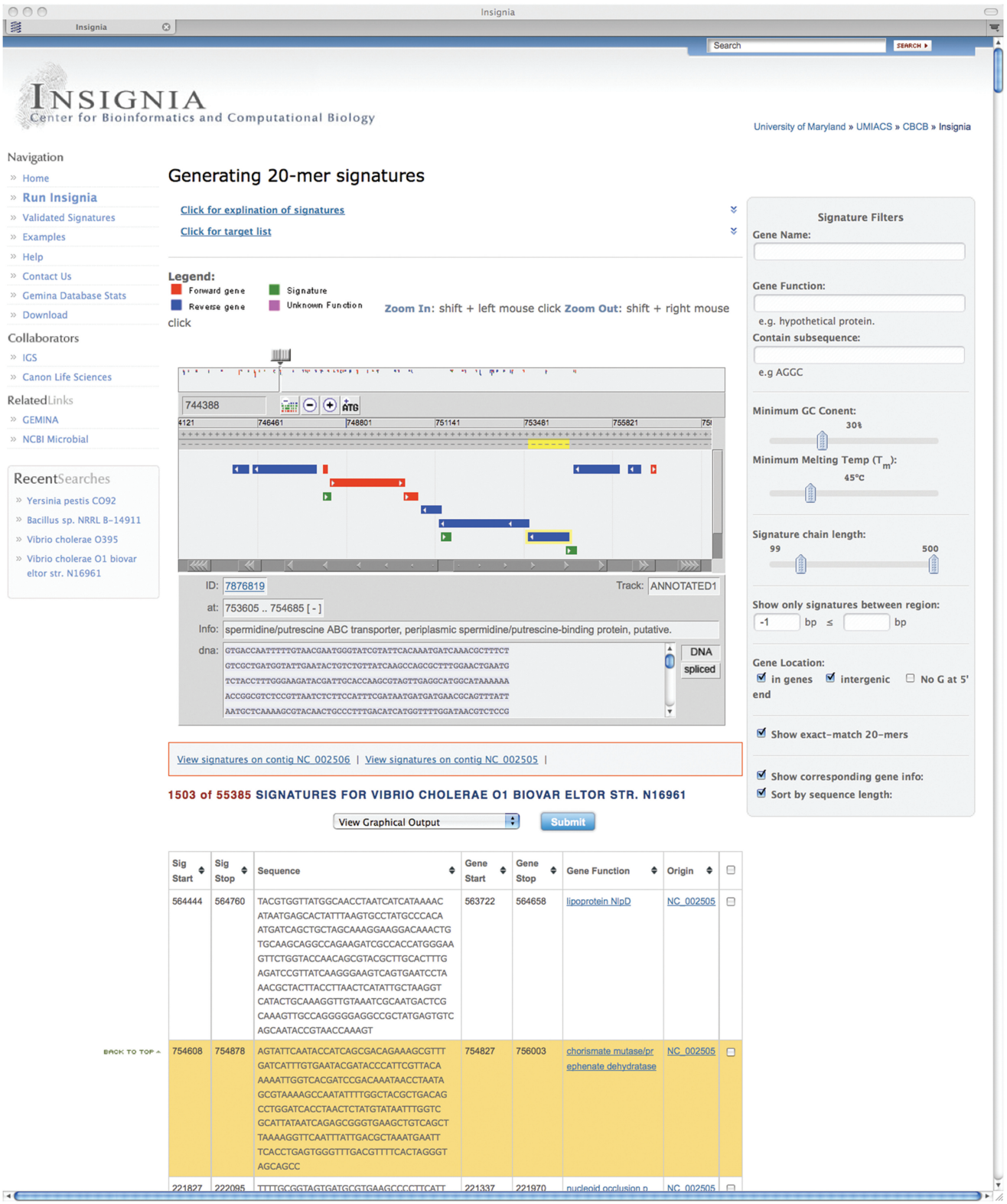

Figure 2. Insignia search results page, showing the graphical genome browser and table of signature sequences. In the genome browser, forward-strand genes are shown in red, reverse-strand genes in blue and signatures in green. The top of the signature table is visible at the bottom of the page, showing start and end positions, signature sequences and annotations. Dynamic signature filters are available in the right margin.

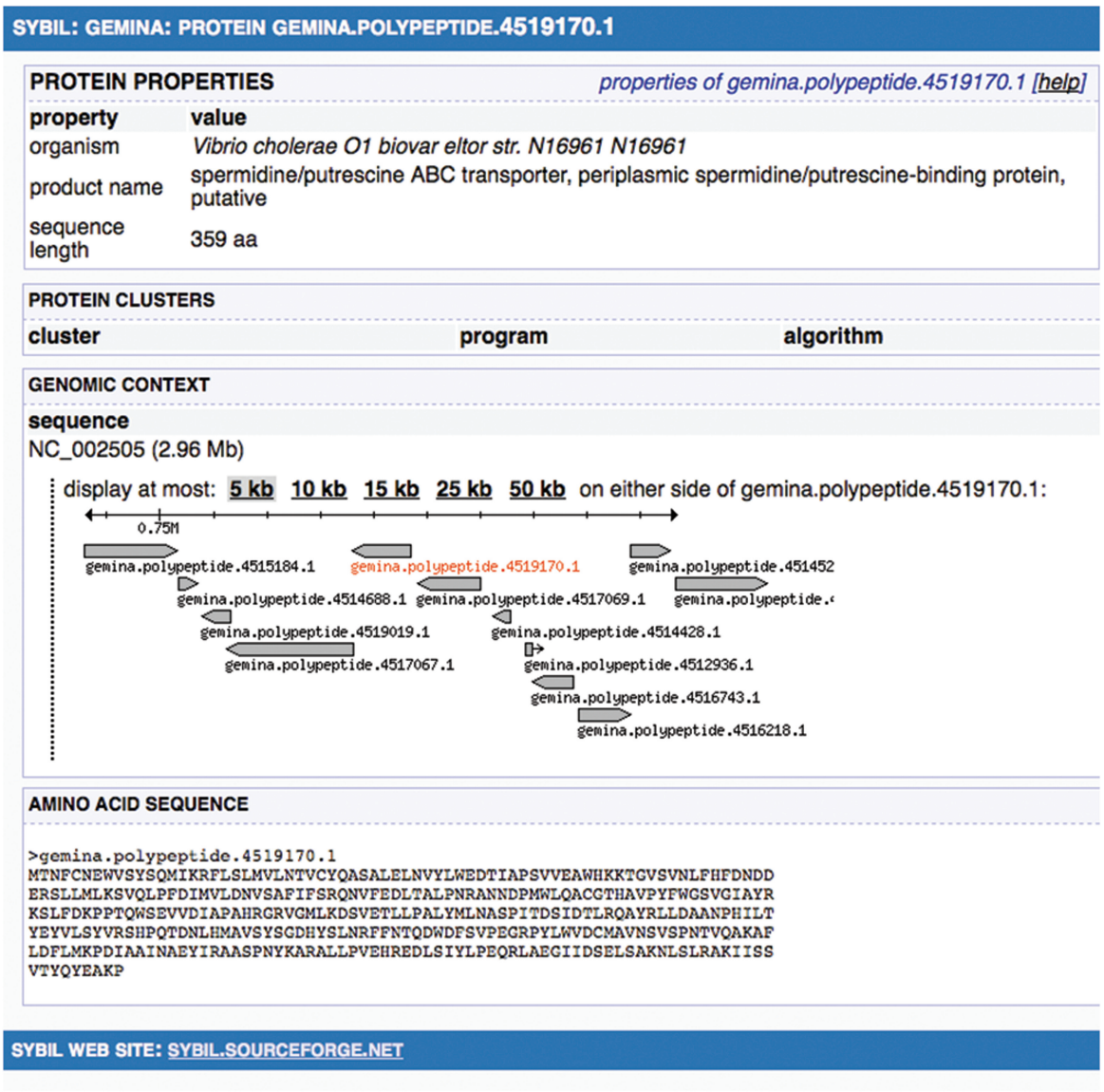

Figure 3. Sybil annotation page displaying a region of the *Vibrio cholerae El Tor N16961* genome. All genes shown in the Insignia genome browser and signature table are linked to a descriptive page describing the gene and its genomic context.

context of their surrounding annotation. In this view, forward-strand genes are shown in red, reverse-strand genes are shown in blue and the signatures are shown in green. Selecting a signature in the genome browser displays its sequence and causes its table row to be highlighted in yellow, as shown at the bottom of the page. Selecting a gene displays its functional annotation and sequence. Further selecting the name or ID of the gene, links to an integrated Sybil page describing the gene product and provides links to the associated resources (Figure 3).

To provide responsive interaction, the signature table and genome browser are limited to displaying at most 2000 signatures. Often, many more signatures are returned from the search and must be filtered to reduce the number. The easiest solution is to gradually increase the length filter until less than 2000 signatures remain. Adjusting any of the filters on the right of the page dynamically updates the signatures displayed in the genome browser and signature table. Signatures can be filtered by gene name, gene function, sequence composition, GC content, melting temperature, length and location. For example, clicking 'Show corresponding gene info' and entering 'toxin' in the 'Gene Function' filter shows only genes with the word 'toxin' in their functional annotation. Performing this filter on the results of the *V. cholerae* query given on the Examples page returns a number of long signatures contained in a gene annotated as 'toxin secretion ATP-binding protein'. With this functional filter, assays designers can target known virulence genes necessary for pathogenicity. Such signatures would be capable of detecting a virulent gene in any genomic context (e.g. within a genetically engineered bacterium).

After filtering the signatures to obtain a manageable number of candidates, users can search the signatures with BLAST and design PCR primers to target the signature. Users can select 'Run BLAST Search' from the Submit menu, to upload the selected signatures to the NCBI BLAST website to perform a more sensitive search and view any alignments to the background. Once the uniqueness is confirmed by BLAST, users can select 'Design Primers' from the Submit menu to design suitable primers for the region using an integrated version of Primer3 (8). Primer3 can be run with default parameters, or with user-defined constraints for assay-specific experimental conditions. TaqMan assays, with a single probe between the primers, can also be designed with the integrated Primer3 software.

CONCLUSION

To date, thousands of Insignia searches have been performed via the web interface. Hundreds of the discovered signatures have been experimentally validated using TaqMan PCR assays for the detection of multiple pathogens, including *V. cholerae* (3), *F. tularensis*, *B. mallei* and *B. pseudomallei* (Cai,M. *et al*. manuscript in preparation). The validation studies have revealed that prioritizing the signature chains by length is an effective strategy, and the validated signatures have shown very little cross-reaction with near-neighbor species. Insignia signatures have also been used for microarray genotyping and detection assays. In all cases, the Insignia signatures were shown to be highly sensitive for detection as well as specific for discrimination between near relatives and environmental backgrounds.

ACKNOWLEDGEMENTS

The authors would like to thank Ivor Knight, Rita Colwell, Mian Cai, Lingxia Jiang, Jacquline Mason, Elisa Taviani, Anwar Huq and Kevin McIver for their supporting work and validation of Insignia signatures; Lynn Schriml and the Institute for Genome Sciences for developing the Gemina database; and the anonymous reviewers for their constructive interface suggestions.

FUNDING

US Department of Homeland Security Science and Technology Directorate under award NBCH2070002 in part. Funding for open access charge: US Department of Homeland Security Science and Technology Directorate under award NBCH2070002.

Conflict of interest statement. None declared.

REFERENCES

1. Slezak,T., Kuczmarski,T., Ott,L., Torres,C., Medeiros,D., Smith,J., Truitt,B., Mulakken,N., Lam,M., Vitalis,E. *et al*. (2003) Comparative genomics tools applied to bioterrorism defence. *Brief Bioinform.*, **4**, 133–149.
2. Vijaya Satya,R., Zavaljevski,N., Kumar,K. and Reifman,J. (2008) A high-throughput pipeline for designing microarray-based pathogen diagnostic assays. *BMC Bioinformatics*, **9**, 185.
3. Phillippy,A.M., Mason,J.A., Ayanbule,K., Sommer,D.D., Taviani,E., Huq,A., Colwell,R.R., Knight,I.T. and Salzberg,S.L. (2007) Comprehensive DNA signature discovery and validation. *PLoS Comput Biol.*, **3**, e98.
4. Lay,M.J. and Wittwer,C.T. (1997) Real-time fluorescence genotyping of factor V Leiden during rapid-cycle PCR. *Clin. Chem.*, **43**, 2262–2267.
5. Mungall,C.J. and Emmert,D.B. (2007) A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics*, **23**, i337–i346.
6. Schriml,L., Gussman,A., Phillippy,K., Angiuoli,S., Hari,K., Goates,A., Jain,R., Davidsen,T., Ganapathy,A., Ghedin,E. *et al*. (2007) Gemina: a Web-Based epidemiology and genomic metadata system designed to identify infectious pathogens. In Zeng,D. (ed.), *BioSurveillance 2007*. Vol. 4506, pp. 228–229.
7. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
8. Rozen,S. and Skaletsky,H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.*, **132**, 365–386.